

REMARKS

Claims 1, 3-4, 6, and 8-14 are pending in the present application. Claim 5 is canceled. Claims 1, 11, and 12 are amended. Claims 2 and 7 were previously canceled in a Response filed on December 21, 2004. Reconsideration of the claims is respectfully requested.

I. 35 U.S.C. § 103, Obviousness, Claims 1, 6, and 10-12

The Examiner has rejected claims 1, 6, and 10-12 under 35 U.S.C. § 103 as being unpatentable over Shriberg et al., incorporated by reference ("Shriberg") in view of Elko et al., U.S. Patent No. 4,741,038 ("Elko"). This rejection is respectfully traversed.

The Examiner bears the burden of establishing a *prima facie* case of obviousness based on the prior art when rejecting claims under 35 U.S.C. § 103. *In re Fritch*, 972 F.2d 1260, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992). For an invention to be *prima facie* obvious, the prior art must teach or suggest all claim limitations. *In re Royka*, 490 F.2d 981, 180 USPQ 580 (CCPA 1974). In this case, the Examiner has not met this burden because all of the features of these claims are not found in the cited references as believed by the Examiner. Therefore, the combination of Shriberg and Elko would not reach the presently claimed invention recited in these claims.

Amended independent claim 1 of the present invention, which is representative of amended independent claims 11 and 12, reads as follows:

1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:
 - determining a fundamental frequency for the digitized audio stream;
 - detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;
extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries;
combining the plurality of prosodic features; and
determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

With regard to claim 1, the Examiner states:

Regarding claims 1 and 11-12, Shriberg et al. disclose a method, a computer usable medium having computer readable program code, and a digital audio processing system for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of: determining a fundamental frequency for the digitized audio stream (*Section 2.1.2.3 on page 133*); detecting changes of the fundamental frequency in the audio stream (*pages 134-135, refer to figure 4*); determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (*pages 134-135*); extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundaries (*section 2.1.1 on page 130 and section 2.1.4 on page 137*); and determining boundaries for the semantic or syntactic units depending on the at least one prosodic feature (*pages 134-135, FO is a prosodic feature*).

Shriberg et al. fail to specifically disclose the step of detecting the changes of the fundamental frequency included providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice. However, Elko et al. teach the step of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice (*col. 11, lines 29-39*).

Since Shriberg et al. and Elko et al. are analogous are because they are from the same field of endeavors, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Shriberg et al.,

by incorporating the teaching of Elko et al. in order to enable the system to pay more coding emphasis on the voice portion than unvoice portion to reduce processing time and increase transmission rate.

Office Action dated September 28, 2005, pages 2-3.

Shriberg teaches prosodic modeling in controlled comparisons for speech data from two corpora: Broadcast News and Switchboard. Shriberg, page 129, section 1.4. More specifically, Shriberg teaches the motivation for each of the prosodic features and specifies the prosodic features extraction, computation, and normalization. Shriberg, page 130, section 1.4. For each inter-word boundary in Shriberg, prosodic features of the word immediately preceding and following the boundary were examined, or alternatively within a window of 20 frames or 200 milliseconds before and after the boundary. Shriberg, page 130, section 2.1.1. In other words, Shriberg teaches the use of a very specific period of time to examine and extract prosodic features in the audio stream. Thus, the only value "empirically optimized" for the method as taught by Shriberg is 200 milliseconds. Shriberg, page 130, section 2.1.1. Figure 1 of Shriberg below further illustrates this teaching:

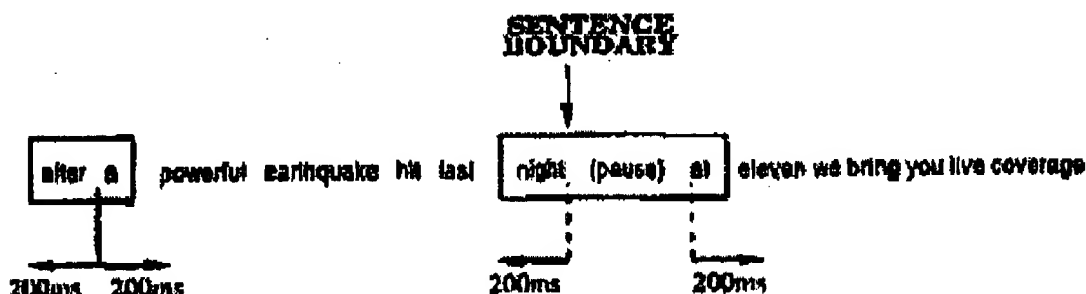


Fig. 1. Feature extraction regions for each inter-word boundary.

As Figure 1 of Shriberg clearly depicts above, the region for feature extraction is 200 milliseconds backward, or to the left, from the pause start and 200 milliseconds forward, or to the right, from the pause end.

In contrast, as amended, the present invention recites in claim 1 extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and

following the candidate boundaries. In other words, when the voicedness of the fundamental frequency estimates are lower than the threshold value, the audio stream contains a voiceless segment creating candidate boundaries and the plurality of prosodic features are extracted in a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries created by the voiceless segment in the audio stream. Support for this amended claim 1 feature may be found in the specification on page 18, line 7 – page 19, line 20 and Figure 4C. By way of example, Figure 4C of the present invention illustrates below that the plurality of prosodic features are extracted from the f1 offset candidate boundary backward, or to the left, for the exemplary period of time of 1000 milliseconds and from the f2 onset candidate boundary forward, or to the right, for another 1000 milliseconds. In addition, Figure 4C of the present invention shows that the voicedness of the fundamental frequency estimates are lower than the threshold value during the 34000-35000+ millisecond time segment.

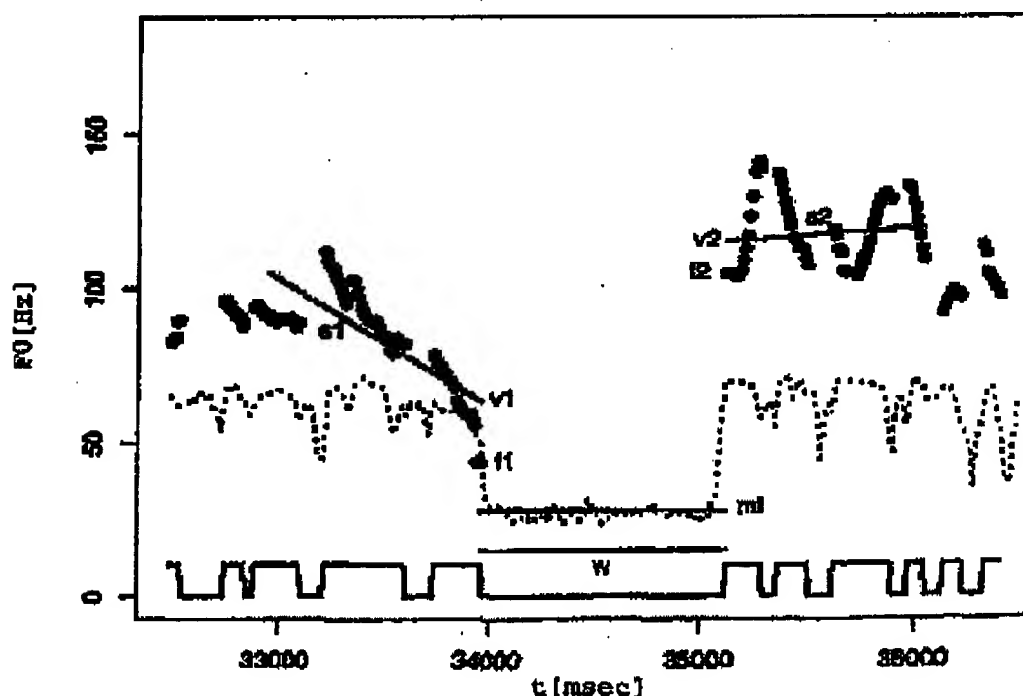


FIG. 4C

Shriberg only teaches one time value of 200 milliseconds for prosodic feature extraction. Amended claim 1 recites a time range between 500 and 4000 milliseconds for

extraction of the plurality of prosodic features. Further, the 200 millisecond time value as taught by Shriberg is not contained within the 500-4000 millisecond time range recited in amended claim 1. Furthermore, it would not have been obvious to one skilled in the art, by applying the teachings of Shriberg, to provide a time range for extraction of prosodic features in order to compare results of varied time intervals for enhanced audio segmentation robustness because Shriberg only teaches one specific time value. Support for this feature may be found in the specification on page 19, lines 16-20. Therefore, Shriberg does not teach or suggest this feature recited in amended claim 1.

Moreover, Shriberg does not teach or suggest combining the plurality of prosodic features and determining boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as further recited in amended claim 1. Shriberg teaches extracting, computing, and normalizing prosodic features. Shriberg, page 130, section 1.4. However, Shriberg makes no reference to combining prosodic features. The focus is on the overall performance, and on analysis of which prosodic features proved most useful for each task. Shriberg, page 130, section 1.4. In determining which prosodic features are most useful, Shriberg analyzes each prosodic feature individually. Shriberg discusses and analyzes each prosodic feature class in individual sections. No section in the Shriberg reference describes prosodic features in combination.

The Examiner cites Shriberg, page 137, section 2.1.4. as teaching the combination of prosodic features. This Examiner-cited section teaches that a feature selection algorithm was utilized to automatically reduce the initial candidate feature set to an optimal subset. Shriberg, page 137, section 2.1.4. Because the initial feature set in Shriberg contained over 100 features, the set is split into smaller subsets. Shriberg, page 137, section 2.1.4. Features are grouped into broad feature classes based on the kinds of measurements involved, and the type of prosodic behavior they are designed to capture. Shriberg, page 131, section 2.1.2. In other words, Shriberg merely places prosodic features into groups or classes depending upon the prosodic feature's behavior or measurements and does not teach or suggest combining the plurality of prosodic features to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as recited in claim 1.

Shriberg does teach however that for each task the results are examined from combining the prosodic information with language model information. Shriberg, page 130, section 1.4. In other words, Shriberg teaches that prosodic information is combined with language model information to evaluate overall performance. Shriberg, page 127, Abstract. But, the combining of prosodic and language models to evaluate performance is not analogous to combining the plurality of prosodic features to determine boundaries for semantic and syntactic units in an audio stream as recited in claim 1. Therefore, Shriberg does not teach or suggest combining the plurality of extracted prosodic features in an audio segmentation process in order to determine semantic or syntactic units as recited in amended claim 1 of the present invention.

Elko does not cure the deficiencies of Shriberg. Elko teaches a method for an acoustic signal processing arrangement that includes at least one directable beam sound receiver adapted to receive sounds from predetermined locations in an environment such as a conference room or an auditorium. Elko, column 2, lines 16-19 and column 3, lines 6-10. "[E]ach of a plurality of directable sound receiving beams receives sound waves from a predetermined location. The sound features signals from the plurality of beams are analyzed to select one or more preferred sound source locations." Elko, column 2, lines 23-27. In other words, the sound signal picked up by each beam is analyzed in a signal processor to form one or more acoustic feature signals and analysis of the feature signals from the different beam directions determines the location of one or more desired sound sources so that a directable beam may be focused thereat. Elko, column 3, lines 10-15. Hence, Elko teaches a method for determining preferred sound source locations in a conference room or auditorium in order to direct sound receiving beams to those preferred sound source locations.

Elko makes no reference to a segmentation process of an audio stream into semantic or syntactic units as recited in amended claim 1. Furthermore, Elko does not teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries and combining the plurality of prosodic features to determine boundaries for the semantic or syntactic units

depending only on the combined plurality of prosodic features as further recited in amended claim 1. Consequently, Elko does not teach or suggest the above recited claim 1 features.

As a result, the combination of Shriberg and Elko does not teach or suggest all limitations recited in amended claim 1 of the present invention. Accordingly, the rejection of independent claims 1, 11, and 12 as being unpatentable over Shriberg in view of Elko has been overcome.

In view of the arguments above, amended independent claims 1, 11, and 12 are in condition for allowance. Claims 6 and 10 are dependent claims depending on independent claim 1. Consequently, claims 6 and 10 also are allowable, at least by virtue of their dependence on an allowable claim. Therefore, the rejection of claims 1, 6, and 10-12 as being unpatentable over Shriberg in view of Elko has been overcome.

II. 35 U.S.C. § 103, Obviousness, Dependent Claims 3-5, 8, and 13

The Examiner has rejected dependent claims 3-5, 8, and 13 under 35 U.S.C. § 103 as being unpatentable over Shriberg in view of Elko, as applied to claim 1, and further in view of Yeldener et al., U.S. Patent No. 5,774,837 ("Yeldener"). This rejection is respectfully traversed.

Dependent claim 5 has been canceled. Consequently, the rejection of claim 5 under 35 U.S.C. § 103 is now moot. However, the language of canceled dependent claim 5 has been incorporated into amended independent claims 1, 11, and 12.

With regard to canceled dependent claim 5, the Examiner states:

Regarding claims 4-5, Shriberg et al. further disclose a method for extracting at least one prosodic feature in an environment of the audio stream where the value of the index function is equal 0 (section 2.1.1 on page 130 discusses feature extraction of both voice and unvoiced portions), that the environment is a time period between 500 and 4000 milliseconds (Section 2.1.1 on page 130).

Office Action dated September 28, 2005, page 5.

As shown in Section I above, Shriberg and Elko do not teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value,

wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in amended independent claims 1, 11, and 12.

Yeldener fails to cure the deficiencies of Shriberg and Elko. Yeldener teaches a method for providing encoding and decoding of speech signals using voicing probability determination. Yeldener, Abstract. More specifically, Yeldener teaches:

...the input speech signal is represented as a sequence of time segments of predetermined length. For each input segment a determination is made as to detect the presence and estimate the frequency of the pitch F_0 of the speech signal within the time segment. Next, on the basis of the estimated pitch is determined the probability that the speech signal within the segment contains voiced speech patterns.

Yeldener, column 4, lines 25-32.

As the passage indicates above, Yeldener teaches that the prosodic feature of pitch is used to determine voiced speech pattern segments. In other words, pitch is the only prosodic feature utilized in the method taught by Yeldener. Consequently, Yeldener cannot teach or suggest extracting or combining a plurality of prosodic features as recited in amended independent claims 1, 11, and 12. Moreover, with regard to the Yeldener reference, the Examiner states:

Regarding applicant's argument in the use of the Yeldener reference, Yeldener is only relied upon for the teaching of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, as agreed by the applicant.

Final Office Action dated April 28, 2005, pages 2 and 3.

As a result, Yeldener does not teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate

boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in amended independent claims 1, 11, and 12. Since Yeldener does not teach or suggest the features recited above in amended independent claims 1, 11, and 12, then the combination of Shriberg, Elko, and Yeldener cannot teach or suggest these independent claim features.

In view of the arguments above, amended independent claims 1, 11, and 12 are in condition for allowance. Claims 3, 4, 8, and 13 are dependent claims depending on independent claims 1 and 12, respectively. Consequently, claims 3, 4, 8, and 13 also are allowable, at least by virtue of their dependence on allowable claims. Accordingly, the rejection of dependent claims 3, 4, 8, and 13 as being unpatentable over Shriberg in view of Elko, as applied to claim 1, and further in view of Yeldener has been overcome.

III. 35 U.S.C. § 103. Obviousness, Dependent Claims 9 and 14

The Examiner has rejected dependent claims 9 and 14 under 35 U.S.C. § 103 as being unpatentable over Shriberg in view of Elko in view of Yeldener, as applied to claims 8 and 13 above, and further in view of Eryilmaz, U.S. Patent No. 5,867,574 ("Eryilmaz"). This rejection is respectfully traversed.

As shown in Section II above, Shriberg, Elko, and Yeldener do not teach or suggest all limitations recited in amended independent claims 1, 11, and 12. In particular, Shriberg, Elko, and Yeldener do not teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in amended independent claims 1, 11, and 12. These recited features are also not taught or suggested in Eryilmaz.

Therefore, because Shriberg, Elko, Yeldener, and Eryilmaz do not teach or suggest extracting a plurality of prosodic features in an environment of the audio stream where the voicedness of the fundamental frequency estimates are lower than the threshold

value, wherein the environment is a period of time between 500 and 4000 milliseconds preceding and following the candidate boundaries; combining the plurality of prosodic features; and determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features as recited in amended independent claims 1, 11, and 12, the combination of Shriberg, Elko, Yeldener, and Eryilmaz cannot teach or suggest these features. As a result, claims 9 and 14 also are allowable at least by virtue of their dependence on allowable claims. Accordingly, the rejection of dependent claims 9 and 14 as being unpatentable over Shriberg in view of Elko in view of Yeldener, as applied to claims 8 and 13 above, and further in view of Eryilmaz has been overcome.

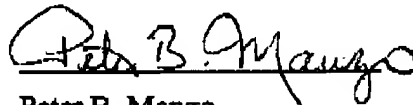
IV. Conclusion

It is respectfully urged that the subject application is patentable over the cited prior art references and is now in condition for allowance.

The Examiner is invited to call the undersigned at the below-listed telephone number if in the opinion of the Examiner such a telephone conference would expedite or aid the prosecution and examination of this application.

DATE: December 22, 2005

Respectfully submitted,



Peter B. Manzo
Reg. No. 54,700
Yee & Associates, P.C.
P.O. Box 802333
Dallas, TX 75380
(972) 385-8777
Attorney for Applicants

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

- ☐ **BLACK BORDERS**
- ☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**
- ☐ **FADED TEXT OR DRAWING**
- ☐ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**
- ☐ **SKEWED/SLANTED IMAGES**
- ☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**
- ☐ **GRAY SCALE DOCUMENTS**
- ☐ **LINES OR MARKS ON ORIGINAL DOCUMENT**
- ☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**
- ☐ **OTHER:** _____

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.